A Machine Learning Framework for Predicting Drug-drug Interactions

Suyu Mei (061021053@fudan.edu.cn)
Shenyang Normal University
Kun Zhang
Xavier University of Louisiana

Research Article

Keywords: drug-drug interaction, drug-target interaction, machine learning, protein-protein interaction networks, signaling pathways

Posted Date: May 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-503867/v1

A machine learning framework for predicting drug-drug interactions

Suyu Mei[*1] and Kun Zhang[*2]

[1]Software College, Shenyang Normal University, Shenyang 110034, China

[2]Bioinformatics Core of Xavier RCMI Center for Cancer Research, Department of Computer Science, Xavier University of Louisiana, New Orleans, LA 70125, US [*]

Corresponding author: Suyu Mei (e-mail: meisygle@gmail.com); Kun Zhang (e-mail: kunzhang@xula.edu).

Abstract

Understanding drug-drug interaction is an essential step to reduce the risk of adverse drug events before clinical drug co-prescription. Existing methods commonly integrate multiple heterogeneous data sources to increase model performance but result in a high model complexity. To elucidate the molecular mechanisms behind drug-drug interactions and reserve rational biological interpretability is a major concern in computational modeling. In this study, we propose a simple representation of drug target profiles to depict drug pairs, based on which an l_2-regularized logistic regression model is built to predict drug-drug interactions. In addition, we develop several statistical metrics to measure the communication intensity, interaction efficacy and action range between two drugs in the context of human protein-protein interaction networks and signaling pathways. Cross validation and independent test show that the simple feature representation via drug target profiles is effective to predict drug-drug interactions and outperforms the existing data integration methods. Statistical results show that two drugs easily interact when they target common genes, or their target genes communicate with each other via short paths in protein-protein interaction networks or through cross-talks between signaling pathways. The unravelled

mechanisms provide biological insights into potential pharmacological risks of known drug-drug interactions and drug target genes.

Keywords drug-drug interaction; drug-target interaction; machine learning; protein-protein interaction networks; signaling pathways

Introduction

Drug-drug interactions (DDIs) have been recognized as a major cause of adverse drug reactions (ADRs) that leads to increasing healthcare costs [1]. DDIs may occur synergistically or antagonistically and potentially result in adverse side effects and toxicities when a patient takes more than one drug concurrently [2]. In many cases, DDIs are hardly detected during clinical trial phase and arbitrary co-prescription of drugs without prior knowledge potentially poses a serious threat to patient health and life [3]. To reduce the risk of potential side effects, we need to check in advance whether two co-prescribed drugs interact. DDIs could be identified via *in vitro*, *in vivo* experiments and *in silico* computational methods. The two former approaches are very costly and are in some cases impossible to be carried out, because DDIs could elicit serious side effects in *in vitro* or *in vivo* experiments [4]. With the advances of pharmacogenomics, recent years have witnessed much progress in developing *in silico* computational methods to predict DDIs and their effects.

Existing computational methods could be roughly classified into three categories, i.e. similarity-based methods [5-8], networks-based methods [9-13] and machine learning methods [14-22]. Similarity-based methods directly calculate similarity scores between drug profiles to infer DDIs. Vilar et al. [5] review several drug profiles that have been used to infer drug repurposing, adverse effects detection and drug–drug interactions, e.g. pharmaceutical profiles, gene expression profiles and phenome profiles. Among

these profiles, drug structural profile could be well interpreted on the assumption that structurally similar drugs tend to target related genes to produce some biological effects [6]. Besides drug profiles, similarity metric is the second major concern in similarity-based methods. Ferdousi et al. [7] compare a dozen of similarity metrics (e.g. inner product, Jaccard similarity, Russell-Rao similarity, Tanimoto coefficient) on drug target profiles and select the optimum measure to infer DDIs. Similarity-based methods have the merit of simple and intuitive interpretation, but the major disadvantage is that these methods cannot resist noise because the thresholding of similarity scores would be seriously affected by potential false DDIs. Networks-based methods could be further divided into two categories, i.e. drug similarity networks-based methods [9-11] and protein-protein interaction (PPI) networks-based methods [12, 13]. The former category of methods first constructs drug-drug similarity networks via a variety of drug similarity metrics and then predicts novel links/DDIs via networks inference, e.g. matrix factorization [9, 10], block coordinate descent optimization [11]. These methods, together with the similarity-based methods [5-8], focus on using drug structural similarities to infer DDIs. Actually, drug-drug interactions often refer to the biological events that two co-prescribed drugs influence or alter each other's therapeutic effects when they take actions on common genes or associated signaling pathways [7]. In this sense, mechanism information is more rational to infer DDIs than drug structural similarity, especially for the prediction of interactions between two structurally dissimilar drugs. Comparatively, the PPI networks-based methods [12, 13] could detect any pair of drugs that simultaneously act on associated genes to produce unexpected effects and thus gain a large coverage of DDIs with less bias. Park et al. [12] assume that DDIs are caused by close interference on common genes or associated genes within the same pathways as well as distant interference through cross-talking pathways, based

on which to capture distant interference via random walk algorithm on PPI networks. Huang et al. [13] also consider drug actions in the context of PPI networks and define the target genes together with their neighboring genes in PPI networks as a target-centred system for each drug, based on which the authors propose a S-score to measure the similarity between two drugs' target-centered systems. PPI networks-based methods have the merit of capturing the mechanism of drug actions but are restricted by the incompleteness of physical PPI networks.

In the last decades, machine learning has attracted increasing attention in inferring drug-drug interactions [14-22]. Most of the machine learning methods focus on data integration to increase the accuracy of DDI prediction. Data integration refers to capturing multiple aspects of a single data source or combining more than one data source. For instance, Dhami et al. [14] explore heterogeneous similarities between drug structures from SMILES representation of drugs (i.e. molecular feature similarity, string similarity, molecular fingerprint similarity and molecular access system), while the other methods focus on integrating multiple data sources. Among the features, drug chemical structural data are the most frequently used [14-21] and are often combined with other data to predict DDIs, e.g. drug adverse drug reactions (ADR) [15-17, 20-21], target similarity [15-17, 19-21], PPI networks [20, 21] and signaling pathways [16]. From methodological point of view, the existing machine learning approaches used to implement data integration majorly include ensemble learning [15, 16], kernel method [14, 17] and deep learning [18, 19]. Data integration is effective to combine multiple data sources or to capture multiple aspects of a single data source. Nevertheless, the performance improvement is often achieved at the cost of model complexity. The existing studies do not reveal which feature information is dispensable and contributes most to DDI prediction and which features are actually less informative. More

100 importantly, data integration imposes intensively demanding constraints on data that
101 are potentially not simultaneously available (e.g. drug and target structures).
102 Comparatively, single data source could reduce model complexity and achieve good
103 interpretability. For instance, Karim et al. [22] use deep neural networks to
104 automatically learn feature representation from the available DDI networks alone to
105 predict novel DDIs. In addition, some inherent mechanisms surely exist behind drug-
106 drug interactions, e.g. common drug-targeted genes and cross-talks between drug-
107 targeted signaling pathways [23]. The underlying DDI mechanisms are often ignored
108 by the existing machine learning methods that focus on drug molecular information
109 itself.

110 In this study, we propose a DDI mechanisms based computational framework
111 without the need of drug structural or ADR similarity information. In this framework,
112 we assume that a drug alters other drugs' therapeutic effects through the associations
113 between their targeted genes or signaling pathways and drug structural similarity is not
114 indispensable for drug-drug interactions. We depict each drug using a drug target
115 profile derived from DrugBank [24]. The profile is actually a binary vector with each
116 element indicating the presence or absence of a gene. The target profiles of two drugs
117 are simply combined as feature vector to depict the drug pair. We extract the known
118 DDIs from DrugBank [24] as positive training data and randomly sample drug pairs as
119 negative training data to train a l_2-regularized logistic regression model. The model is
120 evaluated via cross validation on the training data and independent test on external data
121 from the comprehensive database [25]. Lastly, we further conduct mechanism analyses
122 of the known and the predicted DDIs for further biomedical research.

Data and methods

Data

The known drug-drug interactions (DDI) and drug-target interactions (DTI) are extracted from DrugBank [24]. As we use drug target profile to depict drugs and represent drug pairs, only the drugs that target at least one human gene are studied in this work. As such, we extract 6066 drugs and 2940 human target genes from DrugBank [24]. As results, we totally obtain 915,413 DDIs as the positive training data and obtain 23,169 drug-gene interaction pairs to construct feature vectors representing drug pairs. From the 6066 drugs, we randomly sample 915,413 drug pairs disjoint with the positive training data as the negative training data.

The comprehensive database [25] has curated a large number of DDIs from experiments and text mining. After removing the DDIs that already exist in DrugBank [24], we totally obtain 13 independent datasets, e.g. 8188, 2003 and 37 DDIs from KEGG [26], OSCAR [27] (https://oscar-emr.com/) and VA NDF-RT [28], respectively. These datasets are used as positive independent test data. To estimate the risk of model bias, we randomly sample 8188 drug pairs as negative independent test data, which are disjoint with the training data and the positive independent test data.

To study the underlying molecular mechanisms associated with drug-drug interactions, we construct human physical protein-protein interaction (PPI) networks from HPRD [29], BioGRID [30], IntAct [31] and HitPredict [32]. We totally obtain 171,249 human physical PPIs. In addition, we obtain 27 human immune signaling pathways from NetPath [33], in which IL1~IL11 are merged into one single pathway for simplicity. From Reactome [34], 1846 human signaling pathways are extracted.

Drug target profile-based feature construction

In this study, we investigate drug-drug interactions from the aspect of genes that two drugs act on. For each drug d_i in the DDI-associated drug set D, the human gene set targeted by d_i is assumed to be G_{d_i}. The entire target gene set is defined as follows.

$$G = \cup_{d_i \in D} \ G_{d_i} \quad (1)$$

For each drug d_i, its target profile is formally defined as follows.

$$V_{d_i}[g] = \begin{cases} 1, g \in \ G_{d_i} \wedge g \in G \\ 0, g \notin \ G_{d_i} \wedge g \in G \end{cases} \quad (2)$$

Then the feature vector of a drug pair (d_i, d_j) is represented by combining the target profile of d_i and d_j as follows.

$$V_{(d_i,d_j)}[g] = V_{d_i}[g] + V_{d_j}[g], g \in G \quad (3)$$

The genes $g \notin G$ are discarded. The simple feature representation as described by Formula (3) could intuitively reveal the co-occurrence of common target genes together with unique presence of some target genes between two drugs.

L_2-regularized logistic regression as base learner

We adopt l_2-regularized logistic regression [35] as base learner because it could fast fit large training data and penalize potential overtraining. Given training data *x* and labels *y* consisting of instance-label pairs $(x_i, y_i), i = 1,2,\ldots,l; x_i \in R^n; y_i \in \{-1, +1\}$, the decision function of logistic regression is defined as $f(x) = \frac{1}{1+exp(-y\omega^T x)}$. L_2-regularized logistic regression derives weight vector ω via solving the following prime optimization problem.

$$\min_{\omega} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l} log(1 + e^{-y_i\omega^T x_i}) \quad (4)$$

where C denotes the penalty parameter/regularizer and the second term penalizes potential noise/outlier or overtraining. The optimization problem (4) is solved via its dual form.

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha + \sum_{i:\alpha_i>0}^{l} \alpha_i log\alpha_i + \sum_{i:\alpha_i<C}(C-\alpha_i)log(C-\alpha_i) - \sum_i^l ClogC \qquad (5)$$

$$s.t.\ \ 0 \le \alpha_i \le C, i = 1, \dots, l$$

where α_i denotes Lagrangian operator and $Q_{ij} = y_i y_j x_i^T x_j$. To simplify the parameter tuning, the regularizer *C* as defined in Formula (4) is chosen within the set $\{2^i | -16 \le i \le 16, i \in I\}$, where *I* denotes the integer set.

Model evaluation and DDI mechanism metrics

Binary classification metrics

We adopt five frequently-used performance metrics to evaluate the model performance, i.e. Receiver Operating Characteristic curve AUC (ROC-AUC), sensitivity (SE), precision (PR), Matthews correlation coefficient (MCC), Accuracy and F1 score. ROC-AUC is calculated from the decision values produced by$f(x)$. The other metrics are calculated via a confusion matrix *M,* whose element $M_{i,j}$ records the counts that class *i* are classified to class *j*. From *M*, several intermediate variables are defined first as Formula (6). Then the label-specific PR_l, SE_l and MCC_l for each label are further defined by Formula (7). The overall accuracy and MCC are defined by Formula (8).

$$p_l = M_{l,l}, q_l = \sum_{i=1,i\neq l}^{L} \sum_{j=1,j\neq l}^{L} M_{i,j}, r_l = \sum_{i=1,i\neq l}^{L} M_{i,l}, s_l = \sum_{j=1,j\neq l}^{L} M_{l,j}$$

$$p = \sum_{l=1}^{L} p_l, q = \sum_{l=1}^{L} q_l, r = \sum_{l=1}^{L} r_l, s = \sum_{l=1}^{L} s_l \qquad (6)$$

$$PR_l = \frac{p_l}{p_l + r_l}, l = 1,2\dots,L$$

$$SE_l = \frac{p_l}{p_l + s_l}, l = 1,2\dots,L$$

$$MCC_l = \frac{(p_l q_l - r_l s_l)}{\sqrt{(p_l + r_l)(p_l + s_l)(q_l + r_l)(q_l + s_l)}}, l = 1,2\ldots, L \quad (7)$$

$$Acc = \frac{\sum_{l=1}^{L} M_{l,l}}{\sum_{i=1}^{L}\sum_{j=1}^{L} M_{i,j}}$$

$$MCC = \frac{(pq - rs)}{\sqrt{(p+r)(p+s)(q+r)(q+s)}} \quad (8)$$

where L denotes the number of labels and assumes 2 in this study. F1 score is defined as follows.

$$F1\ \ score = \frac{2 \times PR_l \times SE_l}{PR_l + SE_l}, l = 1\ \ denotes\ \ the\ \ positive\ \ class \quad (9)$$

DDI mechanism metrics

Targeting common genes is one important way that a drug alters other drugs' therapeutic effects to result in drug-drug interactions. We use Jaccard index to measure the intensity that two drugs interact. For a drug pair (d_i, d_j), the Jaccard index between them is defined as follows.

$$Jaccard(d_i, d_j) = \frac{|\ G_{d_i} \cap\ G_{d_j}|}{|\ G_{d_i} \cup\ G_{d_j}|} \quad (10)$$

where G_{d_i} and G_{d_j} denote the target gene set of d_i and d_j, respectively. Given a threshold ξ, we further estimate the percentage of drug pairs whose Jaccard indices exceed ξ as follows.

$$Sim_U = \frac{|\{(d_i,d_j)|Jaccard(d_i,d_j) \geq \xi, (d_i,d_j) \in U\}|}{|U|} \quad (11)$$

where U denotes the set of drug pairs and Sim_U measures the similarities within the set of drug pairs U. If $\xi = min_{\forall (d_i,d_j) \in U} \frac{1}{|\ G_{d_i} \cup\ G_{d_j}|}$, then Sim_U measures the percentage of drug pairs that target at least one common gene.

Two drugs also potentially interact through target genes that communicate via physical paths in human PPI networks. Given a gene pair (g_i, g_j), we use breadth-first graph search algorithm to obtain all the physical paths between g_i and g_j from

human PPI networks, denotes as $P_{(g_i,g_j)}$. The length of the shortest path and longest path within $P_{(g_i,g_j)}$ is denoted as $S_{(g_i,g_j)}$ and $L_{(g_i,g_j)}$, respectively. Then the average number of paths $Avg_{(d_i,d_j)}$, the shortest path length $S_{(d_i,d_j)}$ and the longest path length $L_{(d_i,d_j)}$ between drug d_i and d_j are defined as follows.

$$Avg_{(d_i,d_j)} = \frac{\sum_{(g_i,g_j),g_i \in G_{d_i} \wedge g_j \in G_{d_j}} |P_{(g_i,g_j)}|}{|\{(g_i,g_j)|g_i \in G_{d_i} \wedge g_j \in G_{d_j}\}|}$$

$$S_{(d_i,d_j)} = min_{\forall(g_i,g_j),g_i \in G_{d_i} \wedge g_j \in G_{d_j}} S_{(g_i,g_j)}$$

$$L_{(d_i,d_j)} = max_{\forall(g_i,g_j),g_i \in G_{d_i} \wedge g_j \in G_{d_j}} L_{(g_i,g_j)} \quad (12)$$

These three metrics could measure the communication intensities between two drugs' target genes. Especially, $S_{(d_i,d_j)} = 0$ indicates that there exist common target genes between drug d_i and d_j, and $Avg_{(d_i,d_j)} = 0$ indicates that no paths exist between the target genes of drug d_i and d_j.

Results

Performance of cross validation and independent test

The ROC curve of 5-fold cross validation is illustrated in Figure 1A. The performance is fairly encouraging with ROC-AUC score equal to 0.9877. The other metrics for 5-fold cross validation are provided in Table 1. The metrics of SP, SE and MCC show that the proposed framework shows little bias between the positive and negative classes, e.g. 0.9538 and 0.9394 SE on the positive and negative class, respectively. The overall MMC is up to 0.8983. These results show that simple drug target profile is sufficient to separate interacting drug pairs from non-interacting drug pairs with a high accuracy (Acc=94.66%). Exploration of target genes that two drugs act on helps to clearly unravel molecular mechanisms behind drug-drug interactions and reduce the

information chaos resulting from integration of heterogeneous data. Especially, the information of drug molecule structure is not indispensable and the less informative or irrelevant information would not be introduced into the proposed framework.

To estimate how well the proposed framework generalizes to unseen examples, we further conduct independent test on 13 external DDI datasets and one negative independent test data that are disjoint with the training data. The independent test data sizes are illustrated in Figure 1B varying from 8188 to 3. The recall rates are illustrated in Figure 1C. All the independent test data achieve no lower than 0.8 recall rate except "DDI Corpus 2013". For instance, the proposed framework correctly recognizes 94.97%, 89.92% and 97.30% of the experimental DDIs from KEGG [26], OSCAR [27] and VA NDF-RT [28], respectively (see Table 1). In addition, the proposed framework achieves 0.9373 recall rate on the negative independent test data, indicating a low risk of model bias. The independent test performance shows that the proposed framework trained via simple drug target profile could well generalize to unseen DDIs and meanwhile indicate that the encouraging cross validation performance does not result from overtraining.

Molecular mechanism analyses of drug-drug interactions

Jaccard index between two drugs. The overlap between two drugs' target genes measured via Formula (11) could reveal some common molecular mechanisms that the drugs take effect. The comparison of Jaccard index between the positive training data and the negative training data is illustrated in Figure 1. The threshold of Jaccard index assumes $\xi = min_{\forall(d_i,d_j)\in U} \frac{1}{|\ G_{d_i} \cup\ G_{d_j}|}$ and $\xi = 0.5$ in Figure 2A and Figure 2B, respectively. The results show that interacting drugs tend to target more common genes.

Average number of paths between two drugs. The statistics of paths between two drugs' target genes in PPI networks measured via Formula (12) could reveal the interaction intensity and efficiency between drugs. To reduce the time of paths search, we only randomly choose 9692 interacting drug pairs and 9692 non-interacting drug pairs for molecular mechanism analyses. The average number of paths of top twenty drug pairs are illustrated in Figure 3A. We can see that interacting drug pairs have more paths between target genes than non-interacting drug pairs, indicating that heavy communication traffic between two drugs' target genes tends to cause drug-drug interactions. Additionally, there are fewer interacting drug pairs that no paths exist between their target genes than non-interacting drug pairs as shown in Figure 3B.

Shortest path length between two drugs. Computational results show that the length of shortest paths between two drugs' target genes ranges from 0 to 5. The statistics of shortest path length between two drugs is illustrated in Figure 3C. We can see that interacting drug pairs significantly outnumber non-interacting drug pairs in the case that the shortest path length is equal to 0 (i.e. common target genes). With the increase of shortest path length, non-interacting drug pairs gradually outnumber interacting drug pairs. These results show that common target genes or target genes communicating via shorter shortest physical paths tend to cause drug-drug interactions.

Longest path length between two drugs. Computational results show that the length of longest paths between two drugs' target genes ranges from 0 to 8. As shown in Figure 3D, non-interacting drug pairs outnumber interacting drug pairs when the longest path ranges from 3 to 5, but conversely interacting drug pairs significantly outnumber non-interacting drug pairs when the longest path length equals to 6. These results to some extent show that interacting drugs potentially act on each other via a long range of effect propagation. The metrics $Avg_{(d_i,d_j)}$, $S_{(d_i,d_j)}$ and $L_{(d_i,d_j)}$ as defined in Formula (12)

could measure the tendency of drug-drug interaction from the aspects of interaction intensity, interaction efficacy and action range. The shortest path length equal to 0 and the longest path length equal to 6 are the best indicators to distinguish between interacting drug pairs and non-interacting drug pairs.

Common target pathways between two drugs. We map the target genes onto NetPath [33] and Reactome [34] signaling pathways to gain knowledge about the number of common target pathways between two drugs. Computational results show that interacting drug pairs tend to target more common signaling pathways than non-interacting drug pairs (see Figure 4A for NetPath pathways and Figure 4B for Reactome pathways). Targeting genes located at the same signaling pathways helps to induce synergistic or antagonistic effects of one drug to other drugs.

Common cellular processes between two drugs. Similar to the statistics of common target pathways, the target genes of two interacting drugs are more likely to be involved in common cellular processes than those of two non-interacting drugs (see Figure 4C). This phenomenon is easily interpreted. Two drugs that target genes participating common cellular processes more likely alter each other's therapeutic effects.

Comparison with existing methods

The fundamental difference between this proposed framework and existing methods is that this framework assumes the reactions between target genes of two drugs to be the major driving factor causing drug-drug interactions, while existing methods majorly focus on inferring drug-drug interactions from drug structural similarities. From computational point of view, this framework only uses simple representation of drug target profiles to depict drug pairs so that the model is easy to interpret and the model complexity is under control. However, existing methods integrate multiple heterogeneous data so that we do not know which data is informative and how much

the data contributes to the performance. More importantly, the model complexity is increased and the model is hard to interpret.

From quantitative point of view, we compare this proposed framework with seven existing methods in terms of performance as provided in Table 2. Existing methods generally achieve ROC-AUC scores except Cheng et al. [15] (ROC-AUC=0.67). Nevertheless, these methods generally show imbalanced performance on the positive and negative class. For instance, Vilar et al. [6] use drug structure profile to infer drug-drug interactions but shows extreme bias towards the negative class (e.g. SE 0.68 and 0.96 for the positive and negative class, respectively). Zhang et al. [16] integrate heterogeneous data of drug substructures, drug targets, drug enzymes, drug transporters, drug pathways, drug indications and drug side-effects data, achieving encouraging performance (ROC-AUC score=0.957, PR=0.785, SE=0.670) but only recognizing 7 out of 20 predicted DDIs (equivalent to 35% recall rate of independent test) . Similarly, Gottlieb et al. [20] achieve fairly good performance of cross validation but achieve only 53% recall rate of independent test.

Deep learning has been used to predict effects or types of drug-drug interaction [18, 19]. Karim et al. [22] use deep learning to learn the representation of DDI networks structures and predict novel DDIs. This method also achieves satisfactory performance (ROC-AUC score=0.97, MCC=0.79, F1 score=0.91) but is hard to interpret from the molecular action mechanisms between two drugs. Comparing Table 1 and Table 2, we can see that this proposed framework also outperforms existing methods in terms of performance.

Predictions and clinical implications

We randomly sample 99,986 drug pairs disjoint with the training data and independent test data as prediction set. The proposed framework predicts 43,719 drug pairs to

interact (see Supplementary file S1). These predictions potentially contain a certain level of false interactions. As each prediction is provided with a probability of confidence level, a higher threshold (e.g. 0.7) could be used to filter out those predicted weak interactions. For the predicted DDIs, we further analyse the common cellular processes (see Supplementary file S2) and common signaling pathways (see Supplementary file S3) that the drug target genes are involved in. These analyses help us to understand the underlying mechanisms behind drug-drug interactions. As a case study, we analyse the predicted interaction between drug Nabiximols and Glucosamine from the aspects of common cellular processes and signaling pathways.

Case study on predicted drug-drug interaction (Nabiximols, Glucosamine)

Nabiximols ($C_{42}H_{60}O_4$) extracted from *Cannabis sativa L.* is used for the treatment of neuropathic pain from Multiple Sclerosis and for intractable cancer pain, with the pharmacological effects of analgesic, muscle relaxant, anxiolytic, neuroprotective and anti-psychotic activity (https://www.drugbank.ca/drugs/DB14011). Glucosamine ($C_6H_{13}NO_5$), as a precursor for glycosaminoglycans that are a major component of joint cartilage, is commonly used to rebuild cartilage and treat osteoarthritis (https://www.drugbank.ca/drugs/DB01296). According to DrugBank [24], Nabiximols targets 57 human genes and Glucosamine targets 6 human genes. We further analyse the common cellular processes and signaling pathways that the two drugs' target genes get involved in.

Common cellular processes between Nabiximols and Glucosamine. Computational results show that there are 68 common cellular processes that the target genes of Nabiximols and Glucosamine are involved in. For clarity, we only illustrate 21 cellular processes and their associated target genes in Figure 5 and the rest cellular processes are provided in Supplementary file S2. Two drugs mediate common cellular processes

via common target genes or two associated target genes involved in the same cellular processes. As shown in Figure 5, Nabiximols and Glucosamine mediate the common cellular processes of exogenous drug catabolic process (GO:0042738) and drug metabolic process (GO:0017144) via the common gene *CYP2C19*. In addition, Nabiximols and Glucosamine mediate the common cellular processes of negative regulation of smooth muscle cell proliferation (GO:0048662) via Nabiximols target gene *PPARG* and Glucosamine target gene *IFNG*. Take another example, Nabiximols and Glucosamine mediate the common cellular processes of regulation of reactive oxygen species (ROS) metabolic process (GO:2000377) via Nabiximols target gene *CYP1B1* and Glucosamine target gene *TNF*. Association via different target genes is one major way of two drugs mediating common cellular processes, which potentially leads to synergistic or antagonistic drug-drug interactions. The proposed framework predicts many DDIs that target no common genes but mediate common cellular processes via different target genes (see Supplementary file S2). For instance, drug Nabiximols (DB14011) and Gallium nitrate (DB05260) are not found to target common human genes at present, but they are predicted to target the common cellular processes of neutrophil chemotaxis (GO:0030593), positive regulation of NF-kappaB transcription factor activity (GO:0051092), etc.

Common signaling pathways between Nabiximols and Glucosamine. The common Reactome signaling pathways that Nabiximols and Glucosamine mediate via common or associated target genes are illustrated in Figure 6. Among the target genes, the common target gene CYP2C19 is associated with four Reactome signaling pathways, i.e. Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET) (R-HSA-2142670), Xenobiotics (R-HSA-211981), CYP2E1 reactions (R-HSA-211999) and Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE) (R-HSA-2142816).

Nabiximols and Glucosamine mediate the common signaling pathway of Neutrophil degranulation (R-HSA-6798695) via Nabiximols target gene *ALOX5* and Glucosamine target gene *MMP9*. Mediation of common signaling pathways also potentially leads to drug-drug interactions. Two drugs that do not target common human genes also potentially mediate the same signaling pathways (see Supplementary file S3). For instance, drug Nabiximols (DB14011) and SF1126 (DB05210) have not been reported to target common human genes, but they are predicted to mediate some common signaling pathways, e.g. Regulation of PTEN gene transcription (R-HSA-8943724), Interleukin-4 and Interleukin-13 signaling (R-HSA-6785807), G alpha (q) signaling events (R-HSA-416476).

Discussion

Drug-drug interactions are usually detected and reported only after co-prescribed drugs have clinically done damages to patient health and life. We need resort to computational methods to predict whether two drugs interact before clinical co-prescription to reduce potential risk of side effects. Existing methods generally explore multiple heterogeneous data sources via data integration to increase model performance. The major drawback of these methods lies in the high complexity of model and data. In these methods, we do not know which information contributes to the model performance and interprets the molecular mechanisms behind drug-drug interactions. Among the heterogeneous data, existing methods heavily depend on drug structures with the assumption that structurally similar drugs take into similar therapeutic effects. This assumption surely captures a fraction of drug-drug interactions but shows bias, because it ignores a large fraction of interactions between structurally dissimilar drugs.

Furthermore, data integration would fail when required data are not available, e.g. drug structures, drug side-effects, clinical records.

In this study, we propose a simple representation of drug target profiles to depict drug pairs, based on which to train an l_2-regularized logistic regression model for DDI prediction. As compared to the existing methods, this proposed method could directly interpret the molecular mechanisms behind drug-drug interactions via the association and action between two drugs' target genes in the context of protein-protein interactions (PPI) networks and signaling pathways. We use the known drug-drug interactions from DrugBank as the positive training data and randomly sample the same size drug pairs as the negative training data to train an l_2-regualrized logistic regression model. Computational results show that the proposed framework achieves fairly encouraging performance of cross validation and outperforms the existing methods. Furthermore, the proposed framework demonstrates fairly good performance of independent test on thirteen external DDI datasets. The encouraging performance on the randomly sampled negative independent test data demonstrates no bias towards the positive class. However, the proposed framework demonstrates a little large fraction of false interactions on the prediction set. This result is largely due to the quality of randomly sampled negative training data. Lack of experimentally verified non-interacting drug pairs is a major concern to be addressed in computational modelling. Here we adopt a higher threshold of probability to filter out the weak predictions.

In addition, we develop three statistical metrics to measure the communication intensity, interaction efficacy and action range between two drugs. These metrics help us to understand the statistical characteristic of the paths between two drugs' target genes in human PPI networks. In addition, we use common cellular processes and signaling pathways between two drugs to understand the mechanisms behind drug-drug

interactions. The unravelled mechanisms provide biological insights into potential pharmacological risks of known DDIs and drug target genes.

Conclusions

Integration of heterogeneous data is not indispensable for drug-drug interaction prediction. Simple representation of drug target profile alone achieves encouraging model performance that outperforms existing methods with intuitive elucidation of molecular mechanisms of drug-drug interactions.

Additional Information

File S1 Text file contains the predicted drug-drug interactions.
(TXT)
File S2 Text file contains the genes that are predicted to be targeted by novel drugs for Matador.
(TXT)
File S3 Text file contains the genes that are predicted to be targeted by novel drugs for DrugBank.
(TXT)

Declarations

Statements of ethical approval

Not applicable

Acknowledgement

This work is partly supported by the funding from the NIH grants 2U54MD007595, 5P20GM103424-17 and U19AG055373. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Author Contributions

MS conducted the study and wrote the paper.

Competing interests

The author declares that there no competing interests.

References

1. Wienkers LC, Heath TG. Predicting in vivo drug interactions from in vitro drug discovery data. Nat Rev Drug Discovery 4:825–33 (2005).
2. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 356:1255–9 (2000).
3. Leape LL, Bates DW, Cullen DJ, et al. Systems analysis of adverse drug events. ADE Prevention Study Group. JAMA 274:35–43 (1995).
4. Duke JD, Han X, Wang Z, Subhadarshini A, Karnik SD, et al. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. PLoS Comput Biol 8:e1002614 (2012).
5. Vilar S, Hripcsak G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. Brief Bioinform 18:670-681 (2017).
6. Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, et al. Drug-drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc 19:1066-74 (2012).
7. Ferdousi R, Safdari R, Omidi Y. Computational prediction of drug-drug interactions based on drugs functional similarities. J Biomed Inform 70:54-64 (2017).
8. Vilar S, Uriarte E, Santana L, Lorberbaum T, Hripcsak G et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat Protoc 9: 2147–2163 (2014).
9. Zhang W, Chen Y, Li D, Yue X. Manifold regularized matrix factorization for drug-drug interaction prediction. J Biomed Inform 88:90-97 (2018).
10. Shtar G, Rokach L, Shapira B. Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. PLoS One 14:e0219796 (2019).
11. Zhang P, Wang F, Hu J, Sorrentino R. Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects. Sci Rep 5:12339 (2015).
12. Park K, Kim D, Ha S, Lee D. Predicting Pharmacodynamic Drug-Drug Interactions through Signaling Propagation Interference on Protein-Protein Interaction Networks. PLoS One 10:e0140816 (2015).
13. Huang J, Niu C, Green CD, Yang L, Mei H, Han JD. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. PLoS Comput Biol 9:e1002998 (2013).
14. Dhami DS, Kunapuli G, Das M, Page D, Natarajan S. Drug-Drug Interaction Discovery: Kernel Learning from Heterogeneous Similarities. Smart Health (Amst) 9-10:88-100 (2018).
15. Cheng F, Zhao Z. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc 21:e278-86 (2014).

16. Zhang W, Chen Y3 Liu F, Luo F, Tian G, Li X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics 18:18 (2017).
17. Song D, Chen Y, Min Q, Sun Q, Ye K, Zhou C, et al. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. J Clin Pharm Ther 44:268-275 (2019).
18. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. Proc Natl Acad Sci U S A 115:E4304-E4311 (2018).
19. Lee G, Park C, Ahn J. Novel deep learning model for more accurate prediction of drug-drug interaction effects. BMC Bioinformatics 20:415 (2019).
20. Gottlieb A, Stein GY, Oron Y, Ruppin E, Sharan R. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol 8:592 (2012).
21. Qian S, Liang S, Yu H. Leveraging genetic interactions for adverse drug-drug interaction prediction. PLoS Comput Biol 15:e1007068 (2019).
22. Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network. arXiv:1908.01288 (2019).
23. Jia J, Zhu F, Ma X, Cao Z, Cao ZW, et al. Mechanisms of Drug Combinations: Interaction and Network Perspectives. Nat Rev Drug Discov 8:111-28 (2009).
24. Wishart DS et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074-D1082 (2018).
25. Ayvaz S et al. Toward a Complete Dataset of Drug-Drug Interaction Information From Publicly Available Sources. J Biomed Inform 55：206-17 (2015).
26. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42(Database issue): D199–D205 (2014).
27. Crowther NR, Holbrook AM, Kenwright R, Kenwright M. Drug interactions among commonly used medications. Chart simplifies data from critical literature review. Can Fam Physician 43:1972-6, 1979-81 (1997).
28. Olvey EL, Clauschee S, Malone DC. Comparison of critical drug–drug interaction listings: the department of Veterans Affairs medical system and standard reference compendia. Clin Pharmacol Ther 87:48–51 (2010).
29. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S et al. Human Protein Reference Database--2009 update. Nucleic Acids Res 37(Database issue), D767-72 (2009).
30. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res 43 (Database issue), D470-8 (2015).
31. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. (Database issue) 42, D358–63 (2014).

32. López Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. Database (Oxford) pii: bav117 (2015).
33. Kandasamy K, et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol 11: R3 (2010).
34. Fabregat A, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res46(Database issue): D649–D655 (2018).
35. Fan, R, Chang, K, Hsieh, C, Wang, X & Lin, C. LIBLINEAR: A Library for Large Linear Classification. Mach Learn Res 9, 1871-1874 (2008).

Figure 1 Performance of cross validation and independent test. A. ROC curve and AUC score for 5-fold cross validation. **B.** Statistics of independent test data size. **C.** Recall rates on the independent test data.

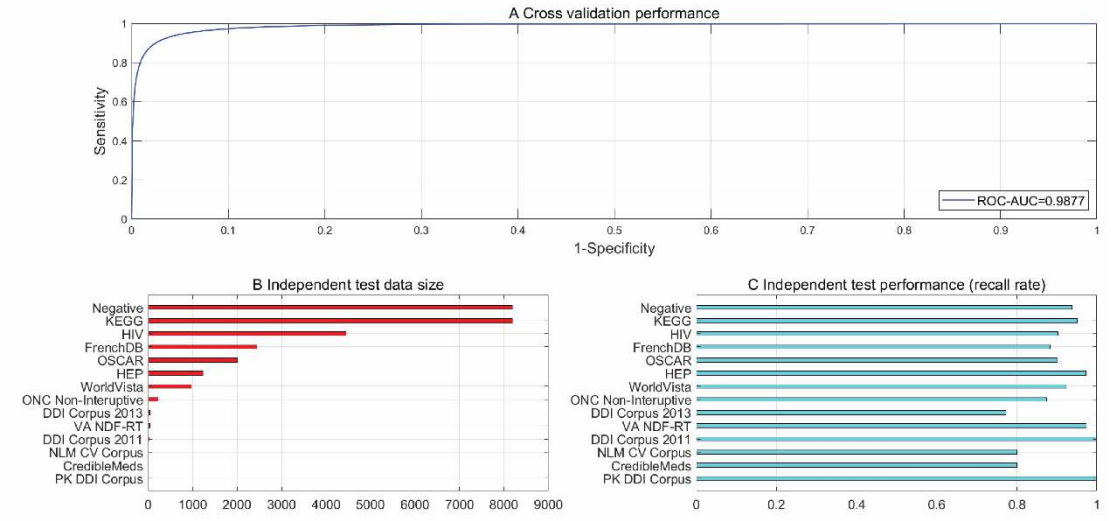

Figure 2 Statistics of common target genes between interacting and non-interacting drugs.

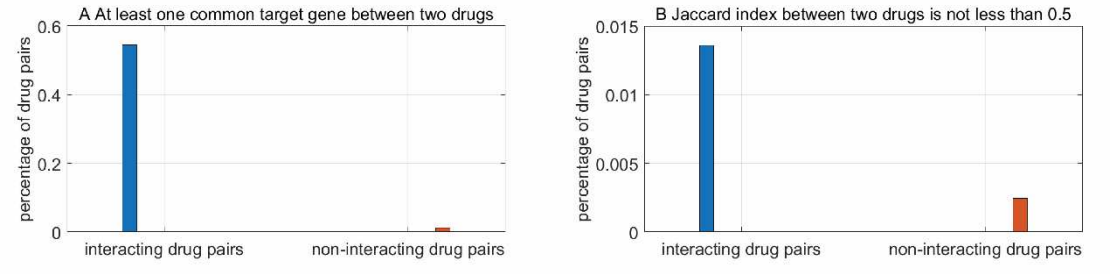

Figure 3 The statistics of average number of paths, shortest path lengths and longest path lengths between two drugs.

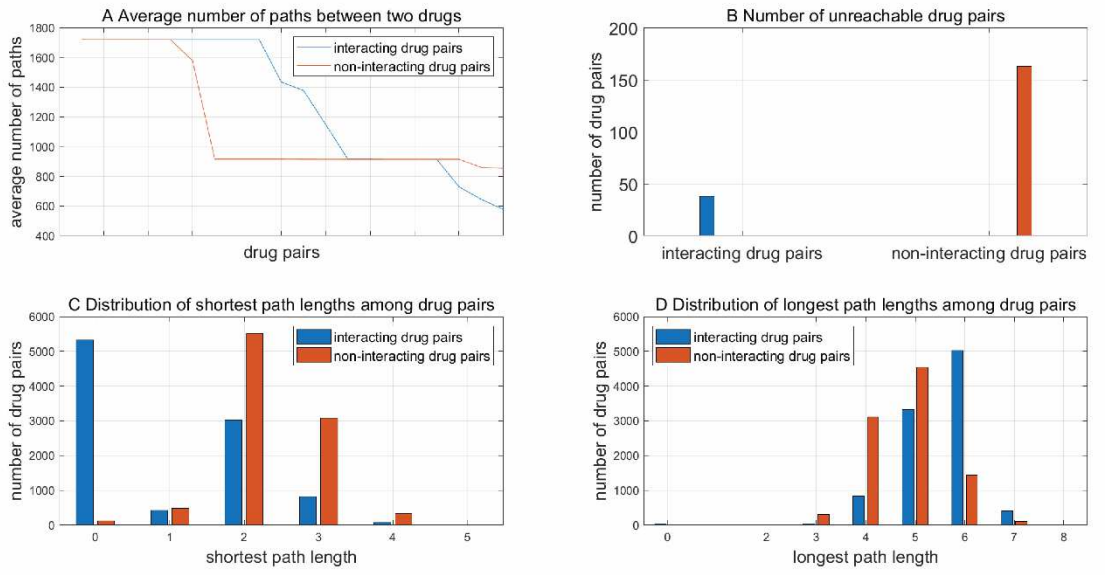

Figure 4 Statistics of common signaling pathways that two drugs target and common cellular processes that two drugs are involved in.

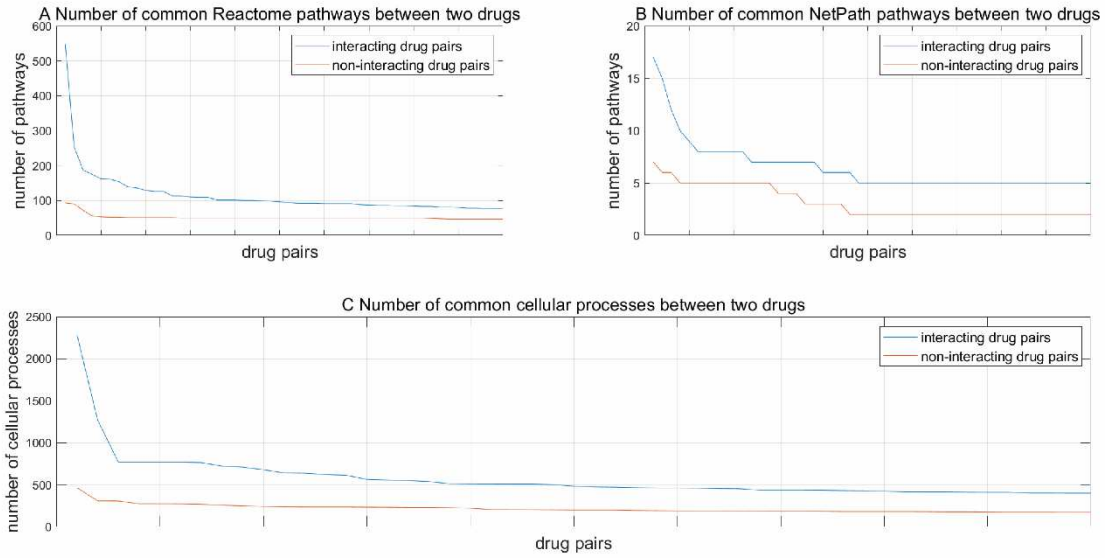

Figure 5 Common cellular processes of target genes between DB14011|Nabiximols and DB01296|Glucosamine predicted to interact. Red triangle nodes denote drugs; green circle nodes denote drug target genes; light red circle nodes denote common target genes; and yellow diamond nodes denote biological processes of gene ontology.

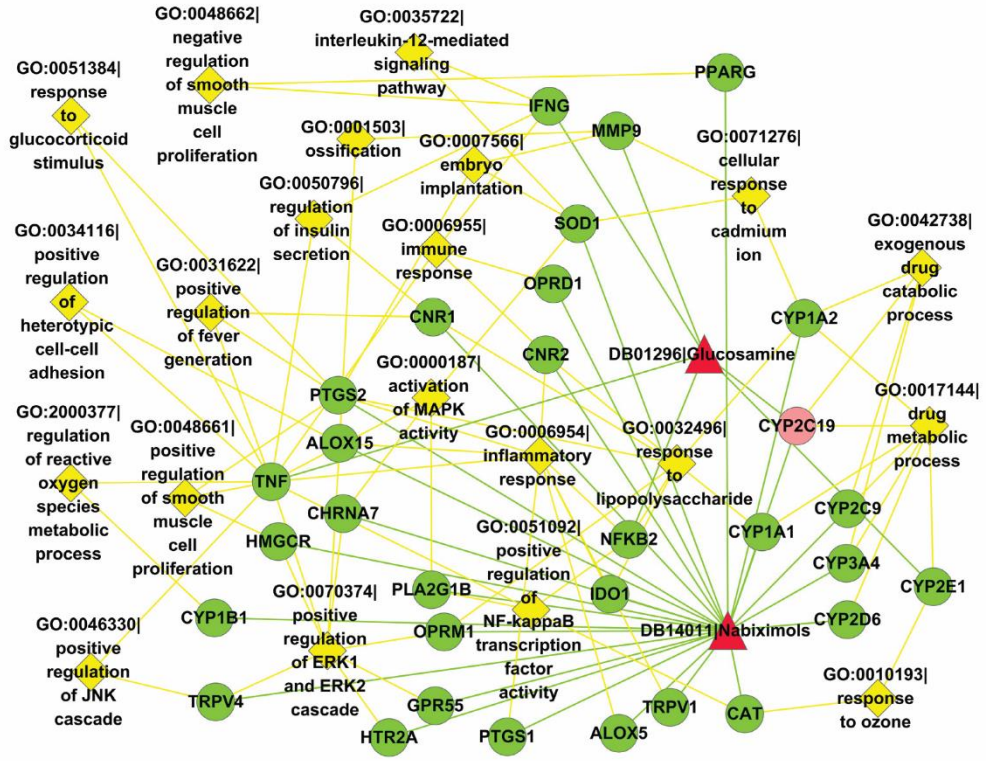

Figure 6 Common target Reactome signaling pathways between DB14011|Nabiximols and DB01296|Glucosamine predicted to interact. Red triangle nodes denote drugs; green circle nodes denote drug target genes; light red circle nodes denote common target genes; and blue hexagon nodes denote Reactome signaling pathways.

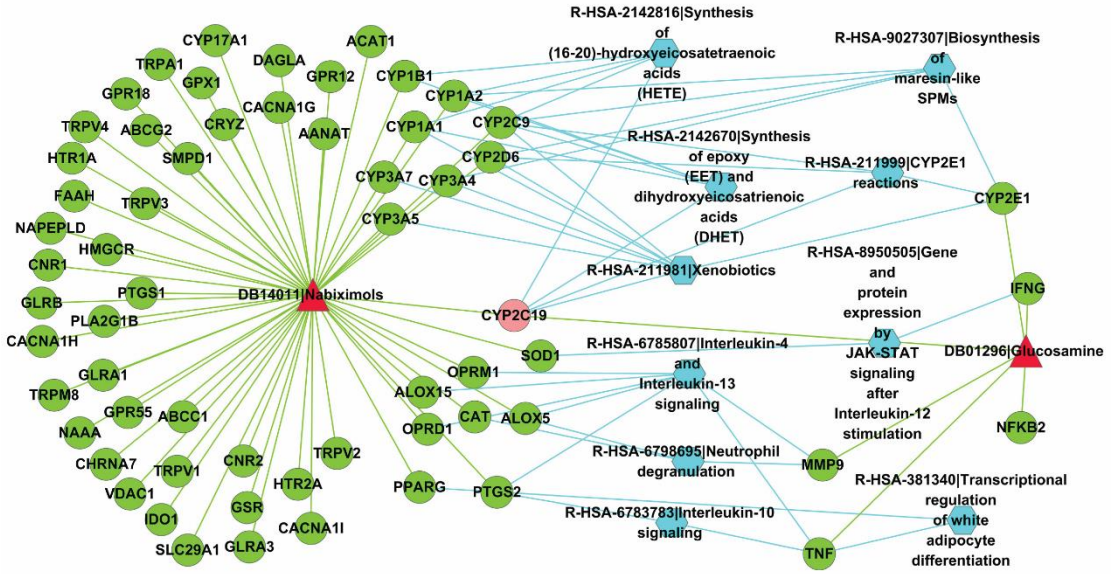

Table 1 Performance estimation of 5-fold cross validation and independent test.

Cross validation							Independent test (recall rate)			
PR	**SE**	**MCC**	**Acc**	**MCC***	**AUC**	**F1 score**	**KEGG**	**OSCAR**	**VA NDF-RT**	**Negative**
0.9402(+) 0.9531(-)	0.9538(+) 0.9394(-)	0.8985(+) 0.8983(-)	94.66%	0.8983	0.9877	0.9470	0.9497	0.8992	0.9730	0.9373

Note: + denotes positive class, - denotes negative class and MCC* denotes overall MCC.

798 **Table 2 Performance comparisons with existing methods.**

	Cross validation					Independent test
	PR	**SE**	**MCC**	**F1 score**	**ROC-AUC**	
Vilar et al. [6]	0.26 (+) 11.81(-)	0.68 (+) 0.96(-)	-	-	0.92	31%
Ferdousi et al. [7]	-	0.72(+)	-	-	-	-
Cheng et al. [15]	-	-	-	-	0.67	-
Zhang et al. [16]	0.785	0.670	-	0.723	0.957	35%
Song et al. [17]	0.68 (+)	-	-	-	0.9738	24%
Gottlieb et al. [20]	0.88	0.93	-	-	0.96	53%
Karim et al. [22]	-	-	0.79	0.91	0.97	-

Note: + denotes positive class and - denotes negative class.

799

Figures

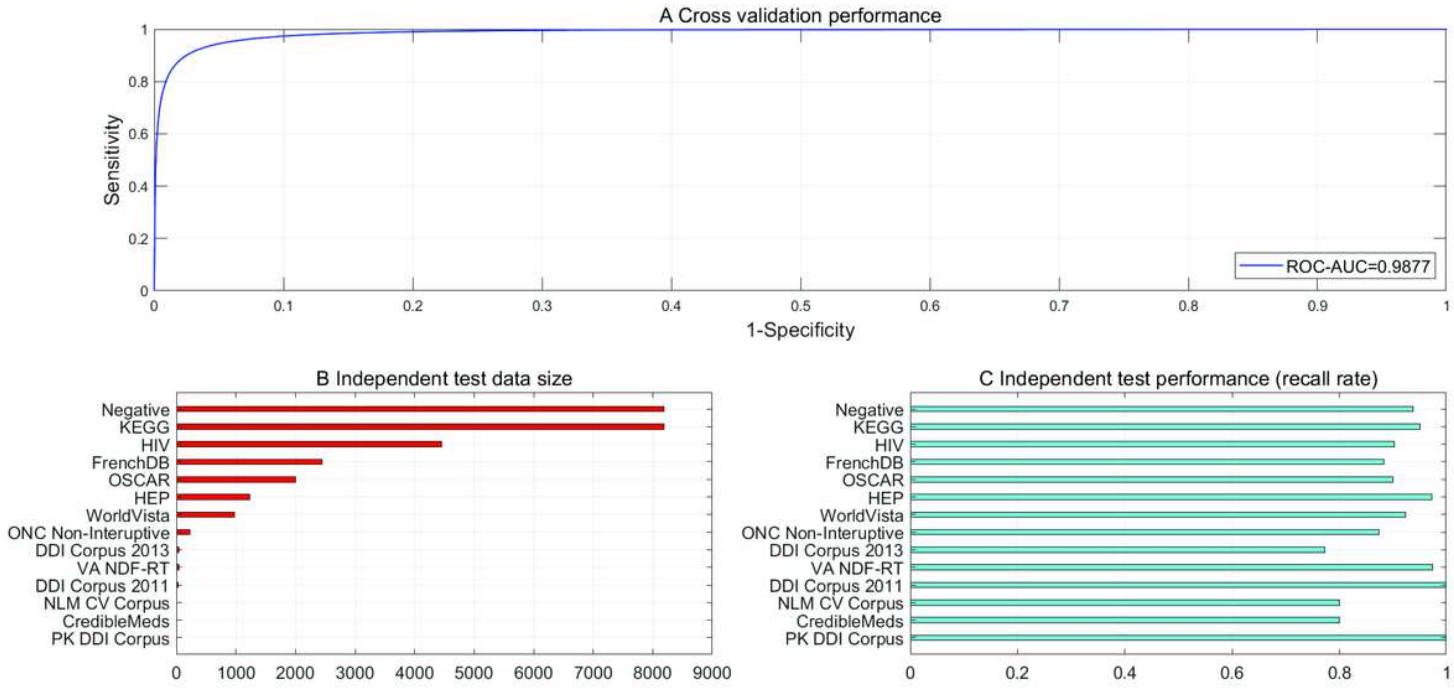

Figure 1

Performance of cross validation and independent test. A. ROC curve and AUC score for 5-fold cross validation. B. Statistics of independent test data size. C. Recall rates on the independent test data.

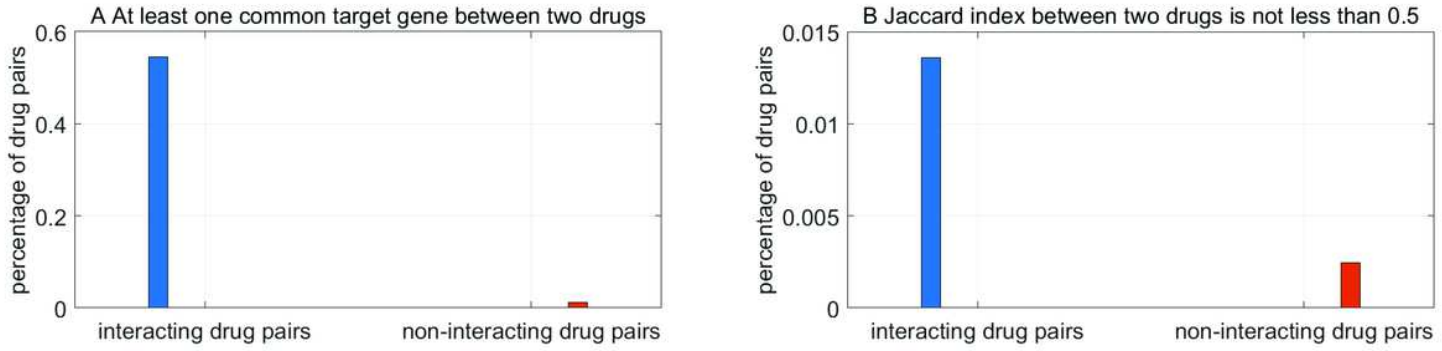

Figure 2

Statistics of common target genes between interacting and non-interacting drugs.

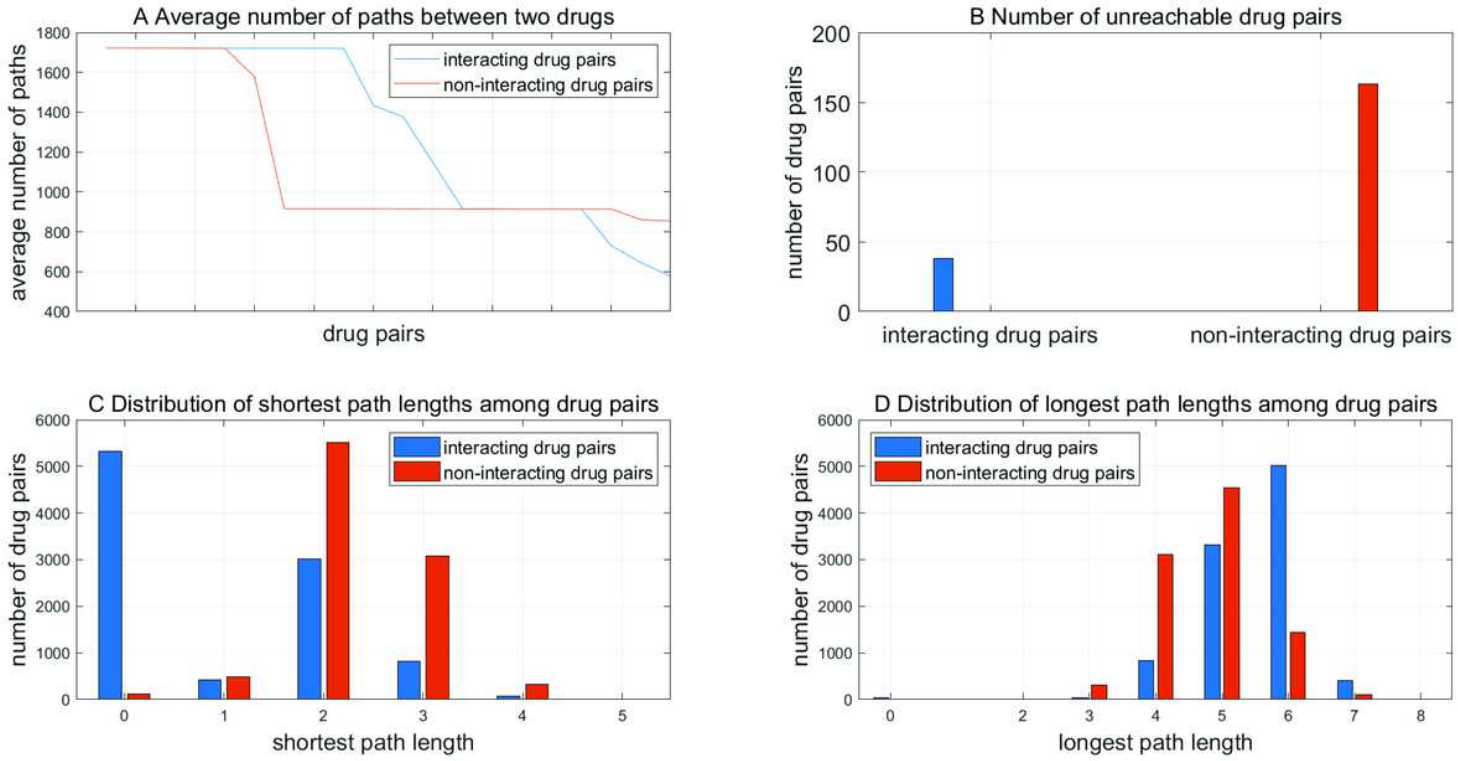

Figure 3

The statistics of average number of paths, shortest path lengths and longest path lengths between two drugs.

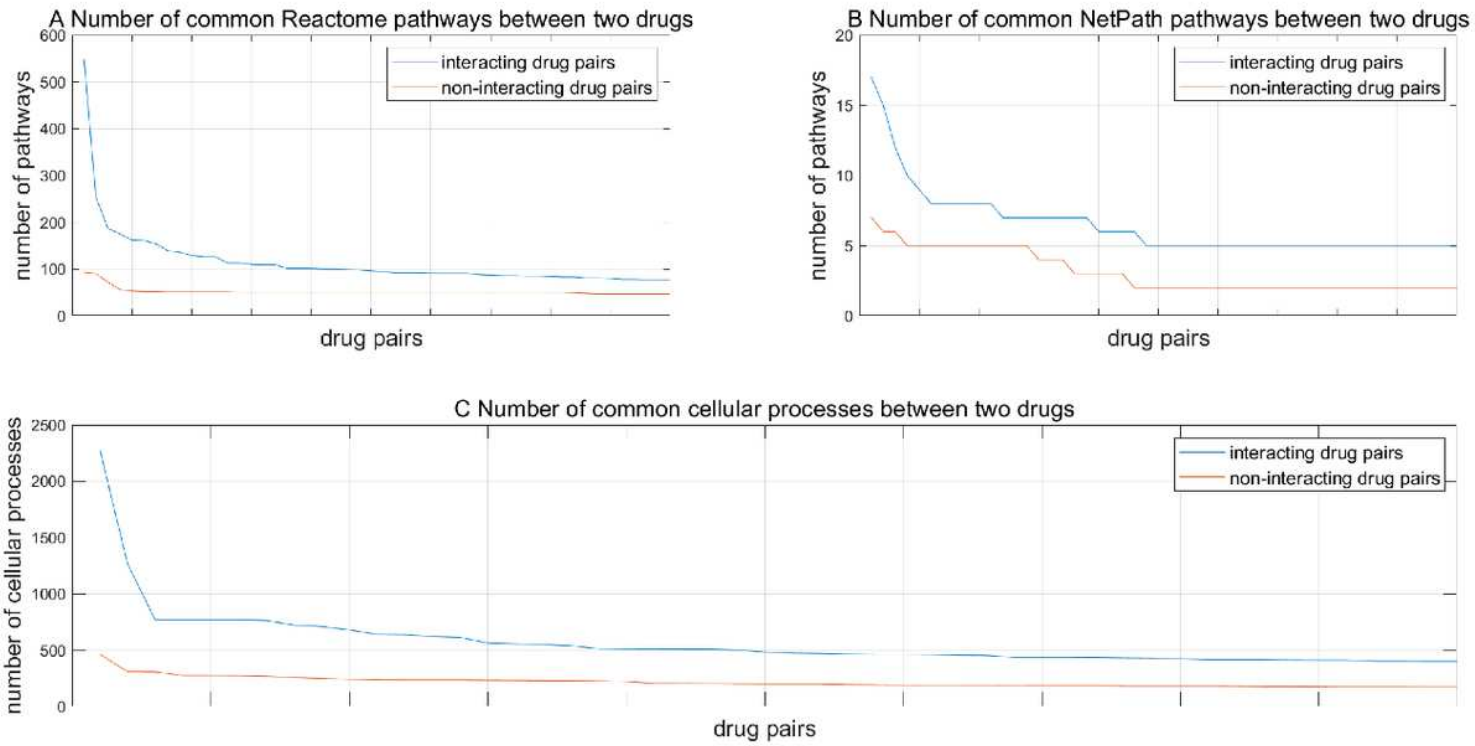

Figure 4

Statistics of common signaling pathways that two drugs target and common cellular processes that two drugs are involved in.

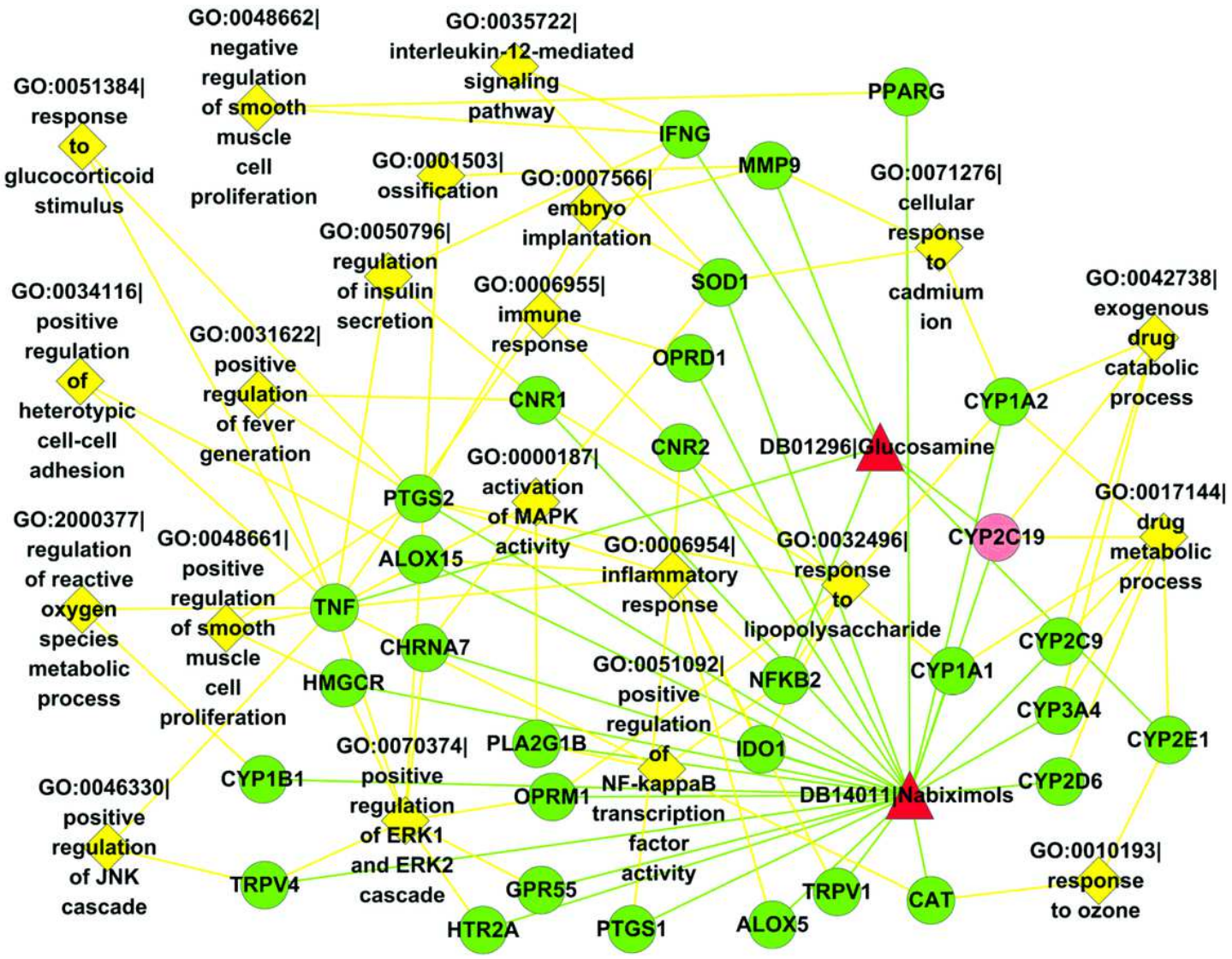

Figure 5

Common cellular processes of target genes between DB14011|Nabiximols and DB01296|Glucosamine predicted to interact. Red triangle nodes denote drugs; green circle nodes denote drug target genes; light red circle nodes denote common target genes; and yellow diamond nodes denote biological processes of gene ontology.

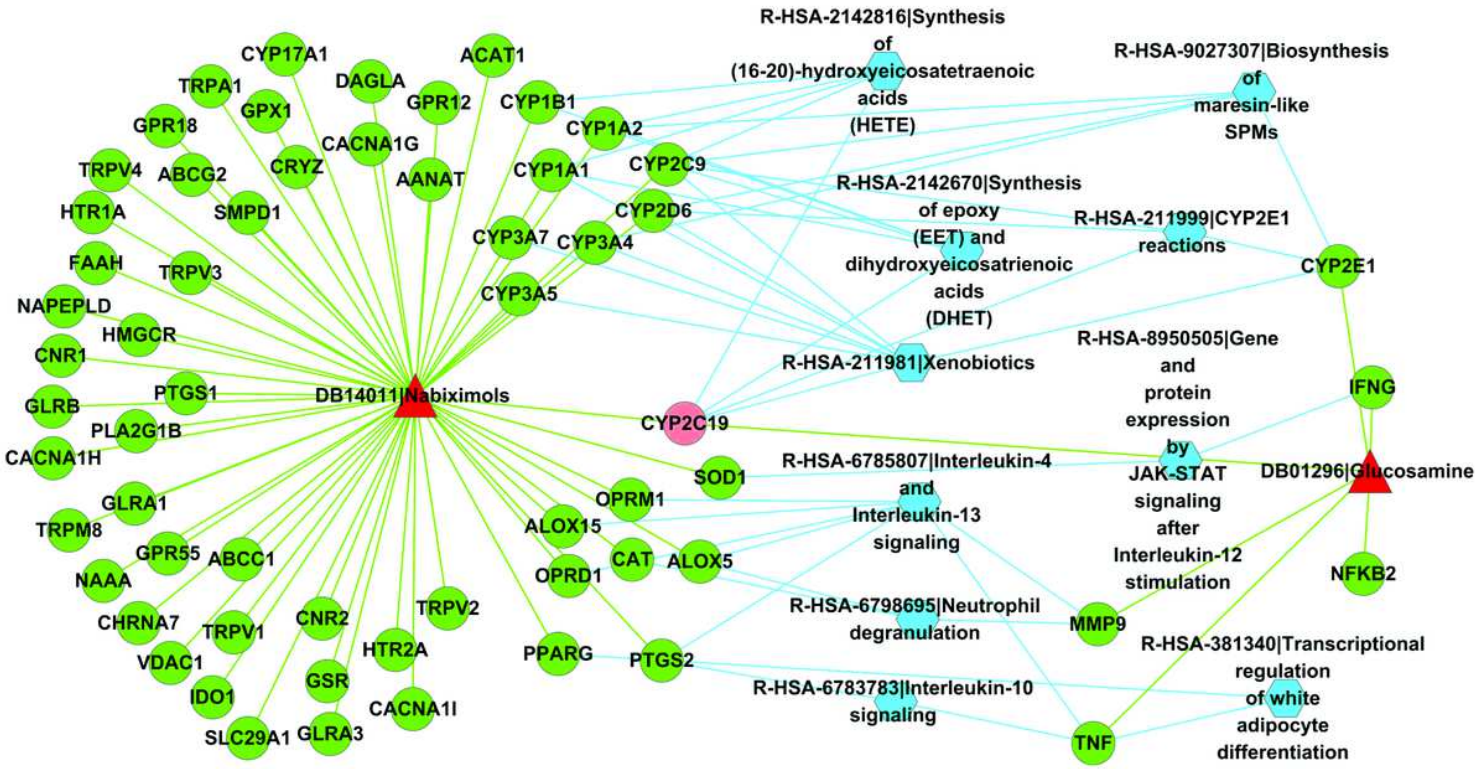

Figure 6

Common target Reactome signaling pathways between DB14011|Nabiximols and DB01296|Glucosamine predicted to interact. Red triangle nodes denote drugs; green circle nodes denote drug target genes; light red circle nodes denote common target genes; and blue hexagon nodes denote Reactome signaling pathways.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1.txt
- S2.txt
- S3.txt